Minimax Estimation in Linear Regression under Restrictions *

Helge Blaker
Department of Mathematics
University of Oslo
Box 1053 Blindern, 0316 Oslo
Norway

March, 1998

Abstract

We consider the problem of estimating the regression coefficients in a linear regression model under ellipsoid constraints on the parameter space. The minimax estimator under weighted squared error is derived. Special cases include ridge regression, Stein's estimator and principal component regression. The asymptotic risk ratio for power ridge regression versus the minimax estimator is computed for a special case and seen to be infinite in some situations. We notice a close connection between this problem and spline smoothing. Adaptive estimators based on Mallows' C_L-statistic are suggested when the size of the parameter space is unknown and shown to have the same asymptotic risk as the estimator based on known size. The minimax estimator is compared to ridge regression and principal component regression on real data sets.

AMS Subject Classification: Primary 62J05; Secondary 62J07, 62F12.
Key words and phrases: Linear regression, minimax estimation, ridge regression, shrinkage, Mallows' C_L, spline smoothing.

*Running head: Minimax regression

1 Motivation

In later years, there has been extensive research devoted to nonparametric regression, largely focused on optimal rates of convergence and bandwidth selection in for instance the kernel method or adaptive selection of other smoothness measures. This article goes the opposite way by using techniques and results from nonparametric regression, in particular spline smoothing, in linear regression. This might also be viewed as an attempt to answer the questions: Why and how should shrinkage be applied in the linear regression model, or what is the proper extension of Stein's estimator to this model?

Under (weighted) squared error loss and with ellipsoid constraints on the regression coefficients, we can compute the minimax estimate of the regression function or the regression coefficients over all linear estimators. The corresponding minimax bound is a special case of the lower bound for minimax mean integrated squared error incurred when estimating the mean of a continuous-time Gaussian process derived by Pinsker (1980). Furthermore, the bound is still attainable asymptotically when the size of the parameter space is unknown and must be estimated.

The minimax linear estimator $\hat{\beta}_M$ considered here is a special case of estimators considered in Pilz (1986) and also has the same form as the minimax spline in Speckman (1985). We compare the minimax linear estimator to ridge and power ridge regression to see how far these estimators are from being asymptotically minimax. This generalizes results from Carter, Eagleson and Silverman (1992) concerning spline smoothers. We also compare the minimax estimator to ridge regression on real data sets, using data driven choices of the shrinkage parameter and show that this adaptive procedure is still asymptotically minimax linear. If the errors are normally distributed, the minimax linear estimator is asymptotically minimax among all procedures. Hence the adaptive procedure, which is completely data-driven, is asymptotically minimax over all estimators in this case.

Estimating the regression coefficients by adaptive minimax regression is thus a practical method with a clear optimality property and a contender to both ridge regression and principal component regression in situations involving multicollinearity where the use of the latter methods is usually advocated. Asymptotic risk calculations show that the gain in risk ratio compared to ridge regression might be infinite for some eigenvalue configurations. The minimax estimator does both shrinkage and variable selection on the principal components and so refines the crude 0-1 shrinkage associated with principal component regression. Computations on real data sets show that the minmax estimator has similar performance to ridge regression but has smaller maximum prediction error.

The paper is organized as follows. Section 2 presents the problem and the technical setup. In section 3, we state a minimax theorem which solves the problem in a special situation while section 4 compares power ridge regression and the minimax estimator in terms of asymptotic maximal risk ratio. In section 5, we make some observations about the similarity between our treatment of the linear regression model and nonparametric regression, in particular spline smoothing. The important problem of adaptive choice of smoothing parameter is addressed in section 6, where it is

shown that estimating the size of the parameter space by Mallows' C_L gives an estimator which is asymptotically minimax. Section 7 compares minimax regression and some reasonable contenders on four well-studied data sets. Proofs are deferred to the appendix.

2 The problem

We consider the familiar regression model

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon} \tag{1}$$

where y is the vector of observations, X is the known design matrix of rank p, β is the vector of unknown regression coefficients and ε is the vector of experimental errors which has mean vector zero and covariance matrix $\sigma^2 I_n$. Let $\mu = X\beta$. The ordinary, or least squares (and MLE if ε is $N(0, \sigma^2 I_n)$) estimator of β is

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y. \tag{2}$$

It should be emphasized that everything is considered in terms of deterministic predictors, so the assumption throughout is that the design matrix X is fixed. Alternatively, the treatment might be conditional on the observed $x-$values and then all conditions on the design must hold a.s. for any sequence $x_1, x_2, \ldots$ of predictors. Large amounts of work deal with improving the estimator (2) with respect to risk or finding more robust estimators with respect to the multicollinearity problem. Some of the techniques developed are ridge regression (Hoerl and Kennard, 1970) and its close relative power ridge regression, variable subset selection in various forms, principal components regression (Massy, 1965), partial least squares, the nonnegative garotte (Breiman, 1995) and the lasso (Tibshirani, 1996). A comparison of all these methods except the lasso and the garotte can be found in Frank and Friedman (1993).

The technical setup is as follows: The loss function is taken to be weighted squared error,

$$L(\hat{\beta}, \beta; A) = (\hat{\beta} - \beta)'A(\hat{\beta} - \beta) \tag{3}$$

where A is an arbitrary positive definite matrix. In particular, $A = X'X$ gives prediction loss $L(\hat{\beta}, \beta; X'X) = ||X\hat{\beta} - X\beta||^2 = ||\hat{\mu} - \mu||^2$. The risk is expected loss, $R(\hat{\beta}, \beta; A) = EL(\hat{\beta}, \beta; A)$. We consider a restricted parameter space of the form, where B is nonnegative definite,

$$\Theta = \{\beta : \beta' B\beta \leq \rho\}. \tag{4}$$

Definition 1 *An estimator β^* is said to be minimax (relative to the parameter space Θ) if*

$$\inf_{\hat{\beta}} \sup_{\beta\in\Theta} EL(\hat{\beta}, \beta; A) = \sup_{\beta\in\Theta} EL(\beta^*, \beta; A)$$

and it is said to be minimax linear if the inf is over all linear estimators, i.e. of form Cy for some matrix C.

It is well known that $\hat{\beta}_{LS}$ is minimax if $\Theta = R^p$.

3 A minimax linear estimator

The following theorem solves the minimax problem in canonical form, which means the problem is rotated into a coordinate system where all matrices are diagonal, thereby greatly simplifying the calculations. Related results can be found in Pinsker (1980) and Pilz (1986).

Theorem 1 *Let* $z_i = d_i\gamma_i + \epsilon_i$, $i = 1, \ldots, p$, *where* $d_i > 0$, $E\epsilon_i = 0$ *and* $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \sigma^2\delta_{ij}$. *Let* $\Gamma = \{\gamma : \sum_{i=1}^p \tilde{b}_i\gamma_i^2 \leq \rho\}$ *be the parameter space and let* $L(\hat{\gamma}, \gamma) = \sum_{i=1}^p \tilde{a}_i(\hat{\gamma}_i - \gamma_i)^2$ *be the loss function where* $\tilde{a}_i$ *and* $\tilde{b}_i$ *are all positive. Let* $\mathcal{C}$ *be the class of all* p *by* p *matrices and set* $\tilde{A} = \mathrm{diag}(\tilde{a}_i)$ *and* $x_+ = \max(x, 0)$. *Then*

$$\begin{aligned}\inf_{C\in\mathcal{C}} \sup_{\gamma\in\Gamma} E(Cz-\gamma)'\tilde{A}(Cz-\gamma) &= \inf_{c_i} \sup_{\gamma\in\Gamma} E\sum_{i=1}^p \tilde{a}_i(c_iz_i-\gamma_i)^2 = \sup_{\gamma\in\Gamma} E\sum_{i=1}^p \tilde{a}_i(c_i^*z_i-\gamma_i)^2 \\ &= \sigma^2\sum_{i=1}^p \tilde{a}_i d_i^{-2}(1-h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+ \qquad (5)\end{aligned}$$

where h *is determined from* $\sum_{i=1}^p \tilde{b}_i\gamma_i^{*2} = \rho$ *where* $\gamma_i^{*2} = d_i^{-2}\sigma^2((\tilde{a}_i/\tilde{b}_i)^{1/2}/h - 1)_+$. *The minimax linear estimator is* $c_i^*z_i$ *where* $c_i^* = (1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+/d_i$.

Let $\tilde{B} = \mathrm{diag}(\tilde{b}_i)$ and $D = \mathrm{diag}(d_i)$. It is easy to see that $\hat{\gamma} = c^*z$ with components $\hat{\gamma}_i = c_i^*z_i$ is the Bayes estimator of γ for the problem in which ϵ and γ are independent normal random vectors, $\epsilon \sim N(0, \sigma^2 I)$ and $\gamma \sim N(0, \Sigma)$ where $\Sigma = \mathrm{diag}(\gamma_i^{*2})$ and h is determined from $\mathrm{tr}(\tilde{B}\Sigma) = \rho$. More precisely, $\hat{\gamma} = E[\gamma|z] = \Sigma(\Sigma+\sigma^2D^{-2})^{-1}D^{-1}z$. This is closely related to the nonparametric minimax Bayes estimator in Heckman and Woodroofe (1991). Now we will reformulate this minimax result in model (1). Let the singular value decomposition of X be $U\tilde{D}V'$, where U is $n \times n$, $\tilde{D}$ is $n \times p$ with elements d_i in position (i, i) where $d_1 \geq d_2 \geq \cdots \geq d_p > 0$ and zero everywhere else, and U and V are orthogonal matrices. Now $X'X = V\tilde{D}'\tilde{D}V' = VD^2V'$ so $d_i = \lambda_i^{1/2}$ where λ_i are the eigenvalues of $X'X$ in decreasing order. Setting $z = U'y$, $\gamma = V'\beta$ and $\epsilon = U'\varepsilon$ transforms (1) to

$$\underset{n\times 1}{z} = \underset{n\times p}{\tilde{D}}\ \underset{p\times 1}{\gamma} + \underset{n\times 1}{\epsilon} \qquad (6)$$

where $E\epsilon = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I_n$, which is covered by Theorem 1 if the loss and prior restrictions transform right. Componentwise,

$$\begin{aligned} z_i &= d_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, p \\ z_i &= \epsilon_i, \qquad\qquad i = p+1, \ldots, n \end{aligned}$$

We need the expressions $(\delta - \beta)'A(\delta - \beta)$ and $\beta'B\beta$ to transform to diagonal forms. A sufficient condition for this is that $X'X$, A and B have the same eigenvectors.

Condition 1 $A = V\tilde{A}V'$ *for* $\tilde{A}$ *diagonal with all elements nonnegative and* $B = V\tilde{B}V'$ *for* $\tilde{B}$ *diagonal with all elements nonnegative.*

Assuming this condition, $\beta' B\beta = \gamma' \tilde{B}\gamma$ and $(\delta-\beta)'A(\delta-\beta) = (\tilde{\delta}-\gamma)'\tilde{A}(\tilde{\delta}-\gamma)$ where $\tilde{\delta} = V'\delta$. Notice for any k, l we have $A^k B^l = V\tilde{A}^k\tilde{B}^k V' = B^l A^k$.

Definition 2 *Let A be a symmetric matrix and let its spectral decomposition be $A = PDP'$ where P is orthogonal and D is diagonal. Define $A_+ = PD_+P'$ where $D_+ = \mathrm{diag}(d_i \vee 0)$.*

With this notation, the minimax estimator in Theorem 1 can be written

$$\hat{\gamma}_M = (I - h\tilde{B}^{1/2}\tilde{A}^{-1/2})_+\hat{\gamma}_{LS} \tag{7}$$

where $\hat{\gamma}_{LS} = (\tilde{D}'\tilde{D})^{-1}\tilde{D}'z$. Theorem 1 then translates to

Corollary 1 *For the model (1), if A and B satisfy Condition 1, the minimax linear estimator is*

$$\hat{\beta}_M = (I - hB^{1/2}A^{-1/2})_+\hat{\beta}_{LS} \tag{8}$$

where h is determined from $\sigma^2\mathrm{tr}\{B(X'X)^{-1}(h^{-1}A^{1/2}B^{-1/2} - I)_+\} = \rho$.

Notice $(I - hB^{1/2}A^{-1/2})_+ = \Sigma^*(\Sigma^* + (X'X)^{-1})^{-1}$ where $\Sigma^* = \sigma^2(X'X)^{-1}(h^{-1}A^{1/2}B^{-1/2} - I)_+$. Let us review some interesting special cases.

1. $A = I$, $B = I$ so $\sum_{i=1}^p \beta_i^2 \le \rho$. The minimax estimator is $\hat{\beta}_M = (1-h)\hat{\beta}_{LS}$ where $h = \sigma^2\sum_{i=1}^p \lambda_i^{-1}/(\sigma^2\sum_{i=1}^p \lambda_i^{-1} + \rho)$ and the minimax risk is

$$\sup_{||\beta||^2 \le \rho} E||\hat{\beta}_M - \beta||^2 = \frac{\rho\sigma^2\sum_{i=1}^p \lambda_i^{-1}}{\sigma^2\sum_{i=1}^p \lambda_i^{-1} + \rho}.$$

2. $A = X'X$, $B = X'X$ gives estimator of the same form, $\hat{\beta}_M = (1-h)\hat{\beta}_{LS}$, but now $h = p\sigma^2/(p\sigma^2 + \rho)$ is different since the loss is $||X\hat{\beta} - X\beta||^2$ and the minimax risk is

$$\sup_{||X\beta||^2 \le \rho} E||X\hat{\beta}_M - X\beta||^2 = \frac{p\sigma^2\rho}{p\sigma^2 + \rho}.$$

This is the natural setting for the Stein estimator considered by Sclove (1968) in regression, $\hat{\beta}_S = (1-\sigma^2(p-2)||X\hat{\beta}_{LS}||^{-2})_+\hat{\beta}_{LS}$. This means estimating the shrinkage factor h by $\min((p-2)\sigma^2/||X\hat{\beta}_{LS}||^2, 1)$. Moreover, if $\rho = \rho^* p$, then for any p by n matrix C,

$$\lim_{p\to\infty}\inf_C \sup_{||X\beta||^2 \le p\rho^*} p^{-1}E||X(CY - \beta)||^2 =$$

$$\lim_{p\to\infty} \sup_{||X\beta||^2 \le p\rho^*} p^{-1}E||X\hat{\beta}_M - X\beta||^2 = \frac{\sigma^2\rho^*}{\sigma^2 + \rho^*}. \tag{9}$$

If $\varepsilon \sim N(0, \sigma^2 I)$, the bound (9) continues to hold if we replace the inf over C by inf over all measurable procedures, i.e. linear estimators are asymptotically minimax, see Pinsker (1980). From Beran (1995), $\hat{\beta}_S$ is asymptotically minimax since for any $\rho^* > 0$,

$$\lim_{p\to\infty} \sup_{||X\beta||^2 \le p\rho^*} p^{-1}E||X\hat{\beta}_S - X\beta||^2 = \frac{\sigma^2\rho^*}{\sigma^2 + \rho^*}.$$

Thus $\hat{\beta}_S$ is an adaptive minimax estimator since it attains the lower minimax bound (9) but does not require knowledge of ρ^*.

3. $A = X'X$, $B = I$ gives $\hat{\beta}_M = (I - h(X'X)^{-1/2})_+\hat{\beta}_{LS}$ where h is found from $\sigma^2 \sum_{i=1}^p \lambda_i^{-1}(\lambda_i^{1/2}/h - 1)_+ = \rho$. This is the natural setting for *ridge regression*. The minimax risk is

$$\sup_{||\beta||^2 \leq \rho} E||X\hat{\beta}_M - X\beta||^2 = \sigma^2 \sum_{i=1}^{p} (1 - h(\lambda_i)^{-1/2})_+.$$

4. $A = X'X$, $B = (X'X)^\delta$ gives $\hat{\beta}_M = (I - h(X'X)^{(\delta-1)/2})_+\hat{\beta}_{LS}$ where h is determined from $\sigma^2 \sum_{i=1}^p \lambda_i^{\delta-1}(\lambda_i^{-\delta/2+1/2}/h - 1)_+ = \rho$ and the minimax risk is

$$\sup_{\beta'(X'X)^\delta\beta \leq \rho} E||X\hat{\beta}_M - X\beta||^2 = \sigma^2 \sum_{i=1}^{p} (1 - h\lambda_i^{(\delta-1)/2}))_+$$

This is the natural setting for *power ridge regression.*

5. $\tilde{B} = \text{diag}(0, \ldots, 0, 1, \ldots, 1)$ with k 0's and $\rho = 0$, i.e. $\sum_{i=k+1}^p \gamma_i^2 = 0$. Then, for any A satisfying condition 1, $c_i^* = I\{i \leq k\}/\sqrt{\lambda_i}$ and $\hat{\gamma}_i = z_i/\sqrt{\gamma_i} I\{i \leq k\}$, which gives $\hat{\beta}_M$ equal to the principal component regression estimator based on the leading k eigenvalues.

6. Let $\tilde{a}_j \neq 0$, $\tilde{a}_i = 0, i \neq j$. Then $c_j^* = (1 + \tilde{b}_j\sigma^2/(\rho\lambda_j))^{-1}{\lambda_j}^{-1/2}$, no restrictions on c_i, $i \neq j$. If this is to hold for all $\tilde{A}$ of this form (i.e. all A with rank 1), the minimax estimator must have this form for all j, or

$$\hat{\beta}_M = (X'X + \sigma^2 B/\rho)^{-1}X'y = (I + \rho^{-1}\sigma^2(X'X)^{-1}B)^{-1}\hat{\beta}_{LS}$$

in original coordinates. This estimator then minimizes $\sup_{\beta'B\beta \leq \rho} E[a'(\hat{\beta}_R - \beta)]^2$ over all vectors a, see Pilz (1986) p.314.

4 A comparison of ridge regression and the minimax linear estimator

A commonly used procedure for estimating β under the assumption $\beta'B\beta \leq \rho$ is to minimize $||y - X\beta||^2$ under the restriction $\beta'B\beta \leq \rho$, leading to the estimator

$$\hat{\beta}_R = (X'X + kB)^{-1}X'y = (I + k(X'X)^{-1}B)^{-1}\hat{\beta}_{LS}. \tag{10}$$

where k is determined from $\hat{\beta}_R'B\hat{\beta}_R = \rho$. If $B = I$ this is ridge regression (Hoerl and Kennard (1970)) and if $B = (X'X)^\delta$ it is power ridge regression. It is interesting to see how much is lost in terms of maximal risk when using ridge og power ridge regression compared to the minimax linear estimator. Clearly, the difference in maximal risk will depend on the eigenvalues and their asymptotic behavior. For instance, if all eigenvalues are equal, both estimators perform constant shrinkage and have the same minimum risk (they are in fact the same procedure). We will restrict attention to $A = X'X$, $B = I$ so $L(\hat{\beta}, \beta) = ||X\hat{\beta} - X\beta||^2$ and the parameter space is $\Theta = \{\beta : \sum_{i=1}^p \beta_i^2 \leq \rho\}$. Under these

assumptions, ridge regression is a natural choice. In canonical coordinates, the ridge and minimax linear estimators are

$$\begin{aligned}\hat{\gamma}_R(k) &= (I + kD^{2(\delta-1)})^{-1}\hat{\gamma}_{LS} \\ \hat{\gamma}_M(h) &= (I - hD^{-1})_+\hat{\gamma}_{LS},\end{aligned}$$

where h and k are the smoothing parameters of the procedures and δ is the power ridge parameter, $\delta = 0$ corresponding to ordinary ridge regression. We know from Theorem 1 that if $\mathcal{L}$ is the class of all linear estimators, then

$$\min_{\hat{\beta}\in\mathcal{L}} \max_{\beta'\beta\leq\rho} EL(\hat{\beta},\beta) = \min_{h} \max_{\beta'\beta\leq\rho} EL(\hat{\beta}_M(h),\beta).$$

For fixed (nonrandom) values of the smoothing parameters, both estimators are linear. In canonical coordinates, the risk for the linear estimator $\hat{\mu}_{L,i} = c_i z_i$ is

$$E||\hat{\mu}_L - \mu||^2 = \sigma^2 \sum_{i=1}^{p} c_i^2 + \sum_{i=1}^{p}(1-c_i)^2\lambda_i\gamma_i^2$$

so

$$\max_{\gamma'\gamma\leq\rho} E||\hat{\mu}_L - \mu||^2 = \rho \max_{1\leq i\leq p}(1-c_i)^2\lambda_i + \sigma^2\sum_{i=1}^{p} c_i^2 \tag{11}$$

Further analysis requires assumptions on the behavior of the eigenvalues to facilitate asymptotic approximations (as $p \to \infty$) of (11). The dimension of the model is thus increasing, and implicitly $p = p(n)$ s.t. $p(n) \to \infty$ when $n \to \infty$. We assume the following holds for the eigenvalues:

$$\lambda_i = C'\frac{p}{i^d} = Ci^{-d} \quad d > 0. \tag{12}$$

where $C = C'p$. This general structure of eigenvalues or special cases have been used in other analyses of linear regression, e.g. Frank and Friedman (1993). While certainly not covering all cases it gives a fairly broad spectrum of different behavior for different values of d.

Proposition 1 *For the minimax linear estimator $\hat{\beta}_M$, we have that the asymptotic minimax risk is*

$$\tilde{R}_M = \left(\frac{\sigma^2 C^{1/d} d}{\rho(d+1)(d+2)}\right)^{d/(d+1)} \rho(d+1) \tag{13}$$

in the sense that

$$\max_{\beta'\beta\leq\rho} E||X\hat{\beta}_M - X\beta||^2 = \tilde{R}_M(1+o(1)) \text{ as } p\to\infty.$$

Proposition 2 *For the power ridge estimator $\hat{\beta}_R$, we have that the asymptotic minimax risk is*

$$\tilde{R}_R = \begin{cases} \left(\frac{\sigma^2 C^{1/d}}{\rho d}\right)^{\frac{d}{d+1}} \rho(d+1) l(\delta,d)^{\frac{1}{d+1}} Int(\delta,d)^{\frac{d}{d+1}} & \delta \leq \frac{1}{2},\ d(1-\delta) > \frac{1}{2} \\ \infty & \delta > \frac{1}{2} \text{ or } d(1-\delta) \leq \frac{1}{2} \end{cases} \tag{14}$$

in the sense that

$$\max_{\beta'\beta \leq \rho} E||X\hat{\beta}_R - X\beta||^2 = \tilde{R}_R(1 + o(1)) \text{ as } p \to \infty.$$

Here,

$$l(\delta, d) = \begin{cases} \frac{(1-2\delta)^{\frac{1-2\delta}{1-\delta}}}{4(1-\delta)^2} & \text{if } \delta < \frac{1}{2} \text{ and } d(1-\delta) > \frac{1}{2} \\ 1 & \text{if } \delta = \frac{1}{2} \text{ and } d > 1. \end{cases}$$

and

$$Int(\delta, d) = \frac{1}{d(1-\delta)}\Gamma\left(\frac{1}{d(1-\delta)}\right)\Gamma\left(2 - \frac{1}{d(1-\delta)}\right) = \frac{1}{d(1-\delta)}\left(1 - \frac{1}{d(1-\delta)}\right) g(\delta, d)$$

where, if $t = \pi/(d(1-\delta))$,

$$g(\delta, d) = \begin{cases} \frac{\pi}{\sin(t)} & d(1-\delta) \geq 1,\ \delta \leq 1/2 \\ \frac{\pi}{\sin(t-\pi)} & 1/2 < d(1-\delta) \leq 1, \delta \leq 1/2. \end{cases}$$

We are now in a position to find the loss of efficiency by using ridge estimators in terms of minimax loss. Combining Propositions 1 and 2, we get

Theorem 2 *The limiting ratio as* $p \to \infty$ *for the minimax risk for power ridge regression compared to the minimax linear estimator is*

$$\frac{\tilde{R}_R}{\tilde{R}_M} = \begin{cases} \infty & \text{if } \delta > \frac{1}{2} \text{ or } d(1-\delta) \leq \frac{1}{2} \\ \left(\frac{(d+1)(d+2)}{d^2}\right)^{\frac{d}{d+1}} l(\delta, d)^{\frac{1}{d+1}} Int(\delta, d)^{\frac{d}{d+1}} & \text{if } d(1-\delta) > \frac{1}{2} \text{ and } \delta \leq \frac{1}{2}. \end{cases}$$

[Figure 1 about here.]

Denote the limit of the risk ratios AMRR(δ, d) (for Asymptotic Maximum Risk Ratio). In particular, for $\delta = 0$ and $d > 1/2$,

$$\text{AMRR}(0, d) = \frac{1}{4}\left(\frac{(d+1)(d+2)}{d^3}\right)^{\frac{d}{d+1}} (4\Gamma(1/d)\Gamma(2 - 1/d))^{\frac{d}{d+1}}$$

It follows from Theorem 2 that

$$\lim_{d\to\infty} \text{AMRR}(\delta, d) = 1 \text{ and } \lim_{d \to 1/(2(1-\delta))} \text{AMRR}(\delta, d) = \infty \tag{15}$$

We can interpret this as follows. Ridge regression, for any $\delta \leq 1/2$, becomes better and better with larger d, i.e. more and more ill-conditioning, since the ratio between largest and smallest eigenvalue is p^d. Notice $\delta = 1$ corresponds to uniform shrinkage. This estimator has infinite minimax risk compared to the minimax estimator for all $d > 0$ (for $d = 0$ they are the same, this is not covered by Theorem 2). Referring to Figure 1, even though the setup is favorable to ordinary ridge regression ($\delta = 0$), it is not entirely clear which choice of δ gives the overall best performance (in the absence of knowledge about $d > 0$). In fact, the less ill-conditioned the problem is, the smaller δ we should choose (negative δ's are allowed), though δ less than about $-1/2$ is not recommended due to large fluctuations in risk ratio.

5 An aside on nonparametric regression

Consider the nonparametric regression model

$$y_i = \mu_i + \varepsilon_i \quad (i = 1, \ldots, n) \tag{16}$$

where $E\varepsilon_i = 0$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{i,j}$, $\mu_i = f(x_i)$, where $0 \leq x_1 < \ldots < x_n \leq 1$ and the unknown function $f \in W_2^k[0,1]$ where $W_2^k[0,1] = \{f : f$ has absolutely continuous derivatives $f', \ldots, f^{(k-1)}$ and $\int_0^1 f^{(k)}(x)^2 dx < \infty\}$. The smoothing spline estimate $\hat{f}_n$ of f is the solution of the optimization problem

$$\min_{f \in W_2^k[0,1]} \sum_{i=1}^{n} (y_i - f(x_i))^2 + h \int_0^1 f^{(k)}(x)^2 dx \tag{17}$$

If we only consider values at the design points, the smoothing spline $\hat{f}_n$ is linear in the y_i's, called the natural polynomial spline of degree $2k-1$ with knots at the x_i's and $f^{(k)} \equiv 0$ on $[0, x_1]$ and $[x_n, 1]$. The choice of smoothness measure enables attention to be concentrated on a finite-dimensional subspace of $C^k[0,1]$, since the solution to (17) is a natural polynomial spline of degree $2k-1$ with knots at the x_i's. Let $\mathcal{S}_n^k$ be the n-dimensional space of natural polynomial splines of degree $2k-1$ with knots at the design points (for details, see Speckman, 1985). There is an orthonormal basis (the Demmler-Reinsch basis) $\{\phi_1, \ldots, \phi_n\}$ for $\mathcal{S}_n^k$ s.t.

$$\sum_{l=1}^{n} \phi_i(x_l)\phi_j(x_l) = \delta_{ij} \text{ and } \int_0^1 \phi_i^{(k)}(x)\phi_j^{(k)}(x)dx = \delta_{i,j}\omega_j \tag{18}$$

where $0 = \omega_1 = \cdots = \omega_k < \omega_{k+1} \leq \cdots \leq \omega_n$. The first k eigenfunctions corresponding to the zero eigenvalues span the space of polynomials of order k. From Speckman (1985), eq. (2.5d),

$$\omega_j = Kn^{-1}(j-k)^{2k}(1 + o(1)) \text{ as } n \to \infty \tag{19}$$

where K is a constant depending on the limiting density of the design points (x_i), e.g. $K = \pi^{2k}$ for uniform design.

Let U_n be the $n \times n$ matrix with element (i,j) equal to $\phi_j(x_i)$. Then $U_n' U_n = I_n$ and $||f^{(k)}||^2 = \sum_{i=1}^n \tilde{f}_i^2 \omega_i$, when $f = \sum_{i=1}^n \tilde{f}_i \phi_i \in \mathcal{S}_n^k$. If we set $\tilde{y} = U_n' y$, $\tilde{f} = U_n' f$ and $\tilde{\varepsilon} = U_n' \varepsilon$, then the model (16) is transformed to

$$\tilde{y}_i = \tilde{f}_i + \tilde{\varepsilon}_i \quad (i = 1 \ldots n), \quad \tilde{f}' \Omega \tilde{f} \leq \rho \tag{20}$$

or $\tilde{y} = \tilde{f} + \tilde{\varepsilon}$ in vector form. Also $||f^{(k)}||^2 = \tilde{f}' \Omega \tilde{f} \leq \rho$ where $\Omega = \mathrm{diag}(\omega_i)$. This is the canonical form of spline smoothing, and it is very close to the canonical form (6) of linear regression. This holds for other bases too provided (18) holds. Clearly the smoothing spline is $\hat{f}_j = (1 + h\omega_j)^{-1}\tilde{y}_j$, while the 'Speckman spline' minimizes $\max\{E||\hat{f} - \tilde{f}||^2; \tilde{f}' \Omega \tilde{f} \leq \rho\}$ over all linear estimators $\hat{f} = c\tilde{y}$, with solution of the form $(1 - h\omega_j^{1/2})_+ \tilde{y}_j$. These procedures are the same as ridge regression and minimax regression, respectively, in canonical form with $\tilde{a}_i = \lambda_i$ and $\tilde{b}_i = \omega_i \lambda_i$. Therefore, theory for spline smoothing is relevant for linear regression. The differences are that the eigenvalues for

the spline smoothing problem are restricted to a particular form whereas they in principle can have any form in linear regression. Also, variance estimation is a lot easier in linear regression while least squares interpolates the data and hence does not display sufficient smoothness in nonparametric regression.

Carter et al.(1992) compare 'Reinsch' and 'Speckman' splines with respect to asymptotic minimax risk when $k = 2$, and their result agrees with $\text{AMRR}(0,4) = (1/4)^{1/5}(45\pi\sqrt{2}/128)^{4/5} = 1.083$ since (19) implies $d = 4$ in Theorem 2 ($\tilde{b}_j = \omega_j\lambda_j = 1$ so $\lambda_j = 1/\omega_j$). In fact, for $d = 4$ this ratio is minimized by $\delta = 0.159$ giving $\text{AMRR}(0.159, 4) = 1.073$ even though the setup is favorable to ordinary ridge regression. But for other eigenvalue combinations, the ratio compares more unfavorably to power ridge regression and might be infinite.

Propositions 1 and 2 can be used to get both rate and constants for minimax risk in nonparametric regression, using (n replaces p now) $C^{1/d} = C'^{1/d}n^{1/d}$ so the rate is $n^{1/(1+d)}$ (or $n^{-d/(1+d)}$ if we normalize risk by n^{-1} which is usually done in rate calculations). If the regression function is k times continuously differentiable, the appropriate d is $2k$. The constant $C' = K^{-1}$ will depend on the density of the design points. In particular, if the design is uniform $C' = \pi^{-2k}$ and Proposition 1 gives

$$\tilde{R}_M = n^{1/(2k+1)}(\rho(2k+1))^{1/(2k+1)}\sigma^{4k/(2k+1)}(k/\pi(k+1))^{2k/(2k+1)} = n^{1/(2k+1)}\gamma(k,\rho,\sigma),$$

say. It then follows that

$$\lim_{n\to\infty}\inf_{\hat{f}}\sup_{f\in W_2^k[0,1]} n^{-1/(2k+1)}E\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2 = \gamma(k,\rho,\sigma)$$

where the inf is over all linear estimators $\hat{f}$. From Pinsker (1980), it follows that the asymptotic minimax risk continues to hold if the minimum is taken with respect to all estimators, provided the ε_i's are independent Gaussian, see also Nussbaum (1985).

Minimax linear estimators constitute an alternative to penalized least squares methods in general. For example, Buja et. al. (1989) discuss linear smoothers as solutions to the penalized least squares problem (where B is a symmetric matrix)

$$||y - f||^2 + hf'Bf$$

where $E[y|x] = f(x)$ and $\text{Var}(y|x) = \sigma^2$. If inverses exist, the solution is $\hat{f} = (I + hB)^{-1}y = Sy$, say. Conversely, if S is an arbitrary symmetric matrix with range $\mathcal{R}(S)$ and $SS^-S = S$, we can obtain $\hat{f} = Sy$ as a stationary solution of

$$Q(f) = ||y - f||^2 + f'(S^- - I)^2 f, \quad f \in \mathcal{R}(S).$$

It can be shown that if $\hat{f} = Sy$ is a symmetric smoother with only non-negative eigenvalues, then $\hat{f}$ is minimax linear over the restricted parameter space $f'(I - S)^2 f \leq \rho$ for some $\rho \geq 0$. For instance, an orthogonal projection S is minimax linear when $||(I - S)f|| = 0$, i.e. $f \in \mathcal{R}(S)$, while a 'constant shrinker' of the form $S = kI$ is minimax under $||f||^2 \leq \rho$ for some $\rho > 0$.

6 Adaptive estimators

The smoothness parameter h in minimax regression is theoretically determined by the size of the parameter space ρ and the variance σ^2, but in practice these are unknown and must be estimated. Alternatively, we can view h at a 'meta-parameter' and select h using some of the procedures used to determine optimal smoothing in curve estimation, e.g. CV, GCV, C_L or other measures. This is parallell to the problem of selecting 'optimal' ridge parameter in ridge regression, e.g. Li (1986). This section describes estimates of ρ and σ which make the corresponding $\hat{\beta}_M$ *asymptotically* minimax among linear estimators (as p and $n-p \to \infty$). Let $\hat{\beta}(h) = C(h)y$ be any linear estimator of β where $C(h)$ is a p by n matrix. Then

$$\begin{aligned} R(Cy,\beta;A) &= E(C(h)y-\beta)'A(C(h)y-\beta) = E[(C(h)y-\hat{\beta}_{LS})'A(C(h)y-\hat{\beta}_{LS})] \\ &\quad +2\sigma^2\mathrm{tr}(A(X'X)^{-1}X'C(h)') - \sigma^2\mathrm{tr}((X'X)^{-1}A) \end{aligned} \tag{21}$$

which gives an unbiased risk estimate if σ^2 is known (or replaced by an unbiased estimate independent of $\hat{\beta}(h)$), see Mallows (1973) p.663. Let

$$\begin{aligned} C_L(h) &= (C(h)y-\hat{\beta}_{LS})'A(C(h)y-\hat{\beta}_{LS}) + 2\sigma^2\mathrm{tr}(A(X'X)^{-1}X'C(h)') \\ &= \sum_{i=1}^{p}\tilde{a}_i(c_i^*(h)z_i - z_i)^2 + 2\sigma^2\sum_{i=1}^{p}\tilde{a}_i c_i^*(h) \end{aligned}$$

where the term independent of h has been dropped and the last line is in terms of the the canonical model (6). The C_L-estimator is the value of h minimizing $C_L(h)$. Li (1986) proved that C_L is asymptotically optimal for selecting the ridge parameter h in (ordinary) ridge regression, i.e. that

$$\frac{||X\hat{\beta}_R(\hat{h}) - X\beta||^2}{\inf_{h\geq 0}||X\hat{\beta}_R(h) - X\beta||^2} \to 1 \text{ as } n \to \infty$$

in probability where $\hat{h} = \mathrm{argmin} C_L(h)$ provided $\inf_h E||X\hat{\beta}_R(h) - X\beta||^2 \to \infty$ as $n \to \infty$. He also proved the same for generalized cross-validation (Golub, Heath and Wahba, 1979) under some additional assumptions on the eigenvalues. Breiman (1995) uses a method he calls 'the little bootstrap' to estimate prediction error. If $A = X'X$, then for procedures like ridge regression and minimax regression, which are based on the same X (i.e. same λ_i) for each value of h, this is the same as C_L. For the minimax estimator $\hat{\beta}_M(h)$, an estimator of h is implicitly an estimator of ρ through Theorem 1. Let $\hat{\gamma}_M(h) = (1-hw_i)_+\hat{\gamma}_{LS}$ and assume $w_i = (\tilde{b}_i/\tilde{a}_i)^{1/2}$ are in increasing order. If the minimizing $\hat{h}$ is in $[w_k^{-1}, w_{k-1}^{-1})$, then $C_L(h)$ is minimized by $\hat{h} = \sigma^2\sum_{i=1}^k \tilde{a}_i^{1/2}\tilde{b}_i^{1/2}\lambda_i^{-1} / \sum_{i=1}^k z_i^2\tilde{b}_i\lambda_i^{-1}$. The corresponding estimator of ρ is

$$\hat{\rho} = \sum_{i=1}^{p}\tilde{b}_i\gamma_i^{*2} = \sum_{i=1}^{k}\tilde{b}_i\lambda_i^{-1}(z_i^2 - \sigma^2).$$

which uses the unbiased estimator $z_i^2/\lambda_i - \sigma^2/\lambda_i$ for γ_i^2 if $i \leq k$ (and 0 for $i > k$).

Kneip (1994) studied large-sample behavior of estimators selected by Mallows' C_L for a large class of estimators he called ordered linear smoothers. It suffices to notice that the minimax linear estimator is an ordered linear smoother. For large p asymptotics we rescale by $\rho = \rho^* p$. The asymptotic minimax risk over linear estimators is

$$\nu^2(A, B, \rho^*, \sigma^2) := \lim_{p\to\infty} \inf_{C} \sup_{\beta' B \beta \le p\rho^*} p^{-1} E(Cy - \beta)' A (Cy - \beta) \tag{22}$$

This is attained by our estimator $\hat{\beta}_M(\hat{h})$ where $\hat{h}$ minimizes $C_L(h)$ and σ^2 is replaced by the standard unbiased estimator.

Theorem 3 *Assume Condition 1 and let* $\{\varepsilon_i\}$ *be iid with* $E[\exp t\varepsilon^2] < \infty$ *for some* $t > 0$ *and assume* $\inf_h E||X\hat{\beta}_M(h) - X\beta||^2 \to \infty$ *when* $p \to \infty$. *Let* $\hat{h} = \text{argmin} C_L(h)$ *where* σ^2 *is estimated by* $\hat{\sigma}^2 = (n-p)^{-1}||y - X\hat{\beta}_{LS}||^2$. *Then, for any* $0 < \rho^* < \infty$, *if* $p \to \infty$ *and* $n - p \to \infty$,

$$\sup_{\beta' B \beta \le p\rho^*} p^{-1} E||X\hat{\beta}_M(\hat{h}) - X\beta||^2 = \nu^2(X'X, B, \rho^*, \sigma^2)(1 + o(1)). \tag{23}$$

Asymptotically, minimax linear risk is attained even though ρ and σ^2 are unknown and estimation of $\hat{\beta}_M$ is completely data-driven. It seems plausible that the same result holds when h is selected by GCV but that remains to be proved. If $\varepsilon_i \sim N(0, \sigma^2)$, $E[\exp t\varepsilon_i^2] = (1 - 2t/\sigma^2)^{-1/2}$ for $0 \le t < \sigma^2/2$. In this case, Pinsker (1980) showed that the bound ν^2 in (22) continues to hold if the inf over C is replaced by inf over all measurable procedures. Therefore, $\hat{\beta}_M(\hat{h})$ is asymptotically minimax over all procedures if the errors are Gaussian.

Example Let $A = B = X'X$ so $\hat{\beta}_M = (1-h)\hat{\beta}_{LS}$ where $h = p\sigma^2/(p\sigma^2 + \rho)$. When ρ and σ^2 are unknown, h is estimated by minimizing $C_L(h) = h^2||X\hat{\beta}_{LS}||^2 + 2(1-h)\hat{\sigma}^2 p$ over $h \in [0, 1]$, giving $\hat{h} = \min(p\hat{\sigma}^2/||X\hat{\beta}_{LS}||^2, 1)$. The adaptive minimax estimator is $\hat{\beta}_M(\hat{h}) = (1 - p\hat{\sigma}^2/||X\hat{\beta}_{LS}||^2)_+ \hat{\beta}_{LS}$ and its asymptotic risk is $\nu^2(X'X, X'X, \rho^*, \sigma^2) = \sigma^2\rho^*/(\sigma^2 + \rho^*)$, compare (9).

The special minimax property (9) enjoyed by $\hat{\beta}_S$ thus generalizes to $\hat{\beta}_M(\hat{h})$ under ellipsoid constraints when $\hat{h}$ is selected by Mallows' C_L. In this sense, adaptive minimax estimation is the proper generalization of Stein's estimator in regression.

7 Practical applications and comparison with other methods

The purpose of this section is to illustrate the predictive performance of $\hat{\beta}_M$ relative to $\hat{\beta}_R$, $\hat{\beta}_{PC}$ and $\hat{\beta}_{LS}$ on some real data sets. We use $A = X'X$ and $B = I$ when computing $\hat{\beta}_M$ and $\hat{\beta}_R$. In light of the assumption $\Theta = \{\beta : \sum_{i=1}^p \beta_i^2 \le \rho\}$, it is natural to rescale to have the design matrix X in correlation form. Certainly the prior assumption is unreasonable if the covariates are on different scales. The smoothing parameters of the different estimators is estimated by minimizing Mallows' C_L.

Each data set is randomly divided into regression and prediction sets of preset sizes, both large. The competing estimators and $\hat{\beta}_{LS}$ are computed from the regression set and the observed squared

prediction error,

$$\mathrm{PSE}(\hat{\beta}) = (1/N_{\mathrm{pred}}) \sum_{\mathrm{pred}} (y_i - \hat{y}_i)^2$$

of each estimator on the prediction set is computed. This is done for a large number (1024) of randomly chosen splits. This way of illustrating predictive performance on real data sets is also found in George and Oman (1996).

7.1 Cement heat evolution data

This data set, taken from Hald (1952) p.647, contains observations of the heat evolved (in calories per gram of cement) on $n = 13$ cement samples of different composition. There are $p = 4$ explanatory variables, giving the amount of different chemicals in the cement mix (in percentage of weight). The data are highly collinear (condition number 1377). A randomized cross-validation was performed, averaging over 1024 random splits with 9 observations in each regression set and the remaining 4 observations used for prediction. Here, $\hat{\beta}_{PC}$ does not perform well and also occationally gives very large observed prediction errors. The estimators $\hat{\beta}_R$ and $\hat{\beta}_M$ improve substantially on $\hat{\beta}_{LS}$ but $\hat{\beta}_R$ sometimes has very large prediction error, i.e. heavy right tail. The minimax estimator $\hat{\beta}_M$ does not suffer from this problem and gives the overall best performance in this example.

Figure 3 shows the "ridge trace" of the estimators, i.e. the components of $\hat{\beta}_M(h)$, $\hat{\beta}_R(h)$ and $\hat{\beta}_{PC}(h)$ as a function of their smoothing parameters h (the constant term is not shown) computed for the full data set. Here, $\hat{\beta}_{PC}(h)$ is the principal component estimator based on the principal components whose corresponding singular values d_i are greater than h. This puts $\hat{\beta}_{PC}(h)$ and $\hat{\beta}_M(h)$ on the same scale. Notice that while $\hat{\beta}_{PC}(h)$ is piecewise constant, $\hat{\beta}_M(h)$ is piecewise linear with knots at the singular values d_i.

[Table 1 about here]

[Figure 2 and figure 3 about here]

7.2 Car price data

The car price data are the file 'auto' described in Becker *et al.* (1988) p.644. The dependent variable is price and there are $p = 11$ independent variables for 74 automobiles. The same adjustments as in George and Oman (1996) were made, i.e. the observations with missing values are eliminated and so is the observation with the highest leverage (nr 73). The dependent variable is converted to logarithms. A randomized cross-validation with 1024 replications, 30 observations in the regression set and the remaining 35 in the prediction set was carried out. The average predictive mean-squared error relative to $\hat{\beta}_{LS}$ can be found in figure 4 and table 2. Here, both $\hat{\beta}_M$ and $\hat{\beta}_R$ outperform $\hat{\beta}_{PC}$, and $\hat{\beta}_M$ has the overall best predictive performance.

[Table 2 about here]

[Figure 4 about here]

7.3 Highway accident data

This data set, analyzed by Weisberg (1980), contains observations of the accident rate for 39 sections of highway in Minnesota in 1973. There are $p = 13$ explanatory variables including some indicator variables. All datapoints are used, and we take 24 observations for each regression set and the remaining 15 for prediction (as in George and Oman, 1996). Figure 5 and table 3 show that $\hat{\beta}_{PC}$ has the best overall performance, giving slightly lower prediction errors than $\hat{\beta}_R$ and $\hat{\beta}_M$ which are similar but with ridge a little better.

[Table 3 about here]

[Figure 5 about here]

7.4 Fish Data

The fish data, taken from Næs (1985), contain observation of 45 samples of rainbow trout. For each sample, fat concentration is determined by ordinary laboratory methods. The $p = 9$ explanatory variables are spectral measurements at different wavelengths from a NIR instrument. The last seven abnormal observations are not used (their also have the seven highest leverage values), see Næs (1985) p.307. The X-matrix is not put in correlation form because the x-variables are already on the same scale. The matrix is extremely collinear, with a conditioning number of $4 \cdot 10^6$. We take 20 observations in the regression set and the remaining 18 in the prediction set. Here, all of the estimators under consideration have significantly lower prediction errors than $\hat{\beta}_{LS}$, with $\hat{\beta}_M$ performing slightly better than $\hat{\beta}_R$. However, $\hat{\beta}_{PC}$ and especially $\hat{\beta}_R$ sometimes have very high prediction errors, while $\hat{\beta}_M$ is never very much worse that $\hat{\beta}_{LS}$.

[Table 4 about here]

[Figure 6 about here]

Proofs

Proof of Theorem 1 First we use an argument from Speckman (1985) p.982 to show that the minimizing matrix C is diagonal. Let $D = \mathrm{diag}(d_i)$.

$$J(C) = \sup_{\gamma \in \Gamma} E(Cz - \gamma)'\tilde{A}(Cz - \gamma) = \sup_{\gamma \in \Gamma} \gamma'(CD - I)'\tilde{A}(CD - I)\gamma + \sigma^2 \mathrm{tr}(\tilde{A}C'C)$$

while

$$J_0(C) = \max_{1 \le i \le p} \rho \tilde{a}_i (c_i d_i - 1)^2 / \tilde{b}_i + \sigma^2 \sum_{i=1}^{n} \tilde{a}_i c_i^2 = J(\mathrm{diag}(C)).$$

Then, if e_j is the j'th unit vector,

$$J(C) \quad \geq \quad \max_{1 \le i \le p} \max_{\tilde{b}_i \gamma_i^2 \le \rho} \gamma_i^2 e_i'(CD - I)'\tilde{A}(CD - I)e_i + \sigma^2 \mathrm{tr}(\tilde{A}C'C)$$

$$= \max_{1\le i\le p} \rho[\tilde{a}_i(c_{ii}d_i-1)^2 + \sum_{j\neq i}\tilde{a}_j c_{ji}^2 d_j^2]/\tilde{b}_i + \sum_{i=1}^p \tilde{a}_i \sum_{j=1}^p c_{ij}^2 \ge J_0(C)$$

with equality if and only if C is diagonal. The minimizing C is thus diagonal. Next, $E(c_i z_i - \gamma_i)^2 = c_i^2\sigma^2 + \gamma_i^2(d_i c_i - 1)^2$ which is mimimized by $c_i = d_i\gamma_i^2/(\sigma^2 + d_i^2\gamma_i^2)$. Thus

$$\sup_\Gamma \inf_{c_i} E\sum_{i=1}^p \tilde{a}_i(c_i z_i - \gamma_i)^2 = \sup_\Gamma \sum_{i=1}^p \tilde{a}_i\sigma^2\gamma_i^2/(\sigma^2 + d_i^2\gamma_i^2) = \nu^2,$$

say. We find ν^2 by the method of Lagrange multipliers. The function

$$\sum_{i=1}^p \tilde{a}_i\sigma^2\gamma_i^2/(\sigma^2 + d_i^2\gamma_i^2) - h^2\sum_{i=1}^p \tilde{b}_i\gamma_i^2$$

is maximized at $\gamma_i^{*2} = d_i^{-2}\sigma^2((\tilde{a}_i/\tilde{b}_i)^{1/2}/h - 1)_+$, where h is determined from $\sum_{i=1}^p \tilde{b}_i\gamma_i^{*2} = \rho$. Then $\nu^2 = \sigma^2\sum_{i=1}^p \tilde{a}_i d_i^{-2}(1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+$. Let $c_i^* = (1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+/d_i$ and observe

$$\sup_\Gamma E\sum_{i=1}^p \tilde{a}_i(c_i^* z_i - \gamma_i)^2 \ge \inf_{c_i}\sup_\Gamma E\sum_{i=1}^p \tilde{a}_i(c_i z_i - \gamma_i)^2 \ge \sup_\Gamma \inf_{c_i} E\sum_{i=1}^p \tilde{a}_i(c_i z_i - \gamma_i)^2 = \nu^2, \tag{24}$$

$$\begin{aligned}
\sup_\Gamma E\sum_{i=1}^p \tilde{a}_i(c_i^* z_i - \gamma_i)^2 &= \sup_\Gamma \sum_{i=1}^p \gamma_i^2(h^2\tilde{b}_i \wedge \tilde{a}_i) + \sum_{i=1}^p \tilde{a}_i\sigma^2 d_i^{-2}(1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+^2 \\
&\le \rho h^2 + \sum_{i=1}^p \tilde{a}_i\sigma^2 d_i^{-2}(1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+^2 \\
&= h^2\sum_{i=1}^p \tilde{b}_i\sigma^2 d_i^{-2}((\tilde{a}_i/\tilde{b}_i)^{1/2}/h - 1)_+ + \sum_{i=1}^p \tilde{a}_i\sigma^2 d_i^{-2}(1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+^2 \\
&= \sigma^2\sum_{i=1}^p \tilde{a}_i d_i^{-2}(1 - h(\tilde{b}_i/\tilde{a}_i)^{1/2})_+ = \nu^2
\end{aligned} \tag{25}$$

so we have equality throughout in (24) and (25).□

Proof of Proposition 1 Let $k = h^{1/2}$ here. Considering (11), $\max_i(1 - c_i)^2\lambda_i = \min(k, \lambda_1)$, recall the eigenvalues are in descending order. But the optimal k is smaller that λ_1, otherwise $\nu^2 = 0$, see (25), so we can restrict attention to $k \le \lambda_1$. Next, let $a = k^{1/2}C^{-1/2}$ and

$$\begin{aligned}
\sum_{i=1}^p c_i^2 &= \sum_{i=1}^p (1 - k^{1/2}\lambda_i^{-1/2})_+^2 \sim \sum_{i=1}^p (1 - ai^{d/2})^2 I(i^d k \le C) \\
&\sim \int_0^{a^{-2/d}} (1 - ax^{d/2})^2 dx = a^{-2/d}\int_0^1 (1 - y^{d/2})^2 dy = \left(\frac{C}{k}\right)^{1/d} \frac{d^2}{(d+1)(d+2)}
\end{aligned}$$

Let

$$T_1(k) = \rho k + \frac{\sigma^2 C^{1/d} d^2}{(d+2)(d+1)} k^{-1/d}$$

so $\tilde{R}_M = \min_k T_1(k)$. The function $f(x) = ax + bx^{-1/d}$ is minimized (for $x \ge 0$) at $x_0 = (ad/b)^{-d/(d+1)}$, with minimum value $f(x_0) = a(d+1)(b/(ad))^{d/(d+1)}$. Hence

$$\tilde{R}_M = \left(\frac{\sigma^2 C^{1/d} d}{\rho(d+1)(d+2)}\right)^{d/(d+1)} \rho(d+1)$$

using $a = \rho$ and $b = \sigma^2 C^{1/d} d^2 (d+1)^{-1}(d+2)^{-1}$. □

Proof of Proposition 2 We have

$$\max_i (1-c_i)^2 \lambda_i = k^2 \max_i \frac{\lambda_i^{2\delta-1}}{(1+k\lambda_i^{\delta-1})^2}$$

Let $f_\delta(x) = x^{2\delta-1}(1+kx^{\delta-1})^{-2}$ for $x \geq 0$. Then

$$\max_x f_\delta(x) = \begin{cases} \frac{1}{4(1-\delta)^2}\left(\frac{k}{1-2\delta}\right)^{(2\delta-1)/(1-\delta)} & \text{when } \delta < 1/2 \\ 1 & \text{when } \delta = 1/2 \\ \infty & \text{when } \delta > 1/2. \end{cases}$$

by straightforward calculus. For $\delta < 1/2$, the function is maximized at $x = (k/(1-2\delta))^{1/(1-\delta)}$. Therefore, since i^d/p becomes dense in $[0, K]$ for any finite K as $p \to \infty$ for fixed $d > 0$, $i = 1, \ldots, p$, $\max_i (1-c_i)^2\lambda_i = \infty$ when $\delta > 1/2$ and otherwise, as $p \to \infty$,

$$\max_i (1-c_i)^2 \lambda_i \sim \begin{cases} k^2 & \text{when } \delta = 1/2. \\ \frac{k^{1/(1-\delta)}}{4(1-\delta)^2}(1-2\delta)^{(1-2\delta)/(1-\delta)} & \text{when } \delta < 1/2. \end{cases}$$

Next, let $a = kC^{\delta-1}$ and compute

$$\begin{aligned} \sum_{i=1}^p c_i^2 &= \sum_{i=1}^p (1+k\lambda_i^{\delta-1})^{-2} = \sum_{i=1}^p (1+ai^{d(1-\delta)})^{-2} \sim \int_0^\infty (1+ax^{d(1-\delta)})^{-2} dx \\ &= a^{-1/(d(1-\delta))} \int_0^\infty (1+y^{d(1-\delta)})^{-2} dy = C^{1/d} k^{-1/(d(1-\delta))} Int(\delta, d) \end{aligned}$$

when $d(1-\delta) > 1/2$ and infinity otherwise. Let

$$T_2(k) = \rho \frac{k^{1/(1-\delta)}}{4(1-\delta)^2}(1-2\delta)^{(1-2\delta)/(1-\delta)} + \sigma^2 C^{1/d} k^{-1/(d(1-\delta))} Int(\delta, d)$$

for $\delta < 1/2$, $d(1-\delta) > 1/2$. The special case for $\delta = 1/2$ is defined by continuity. Let $g(x) = ax^{1/(1-\delta)} + bx^{-1/(d(1-\delta))}$ which is maximized at $x_0 = (b/(ad))^{d(1-\delta)/(d+1)}$ with maximum $g(x_0) = (b/(ad))^{d/(d+1)} a(d+1)$. This gives the maximum risk formula when $a = \rho(1-\delta)^{-2}(1-2\delta)^{(1-2\delta)/(1-\delta)}/4$ and $b = \sigma^2 C^{1/d} Int(\delta, d)$. □

Proof of Theorem 3 Change to canonical form and let $\mu_i = \sqrt{\lambda_i}\gamma_i$ and $\hat{\mu}_{i,M} = \hat{\gamma}_{i,M}\sqrt{\lambda_i}$. Recall the parameter space is $\Theta = p^{-1}\sum_{i=1}^p \mu_i^2 \tilde{b}_i/\lambda_i \leq \rho^*$. Eq.(1.5) in Kneip (1994) gives

$$\sup_{\mu\in\Theta}((Ep^{-1}\sum_{i=1}^p(\hat{\mu}_{i,M}(\hat{h}) - \mu_i)^2)^{1/2} - (\inf_h Ep^{-1}\sum_{i=1}^p(\hat{\mu}_{i,M}(h)_i - \mu_i)^2)^{1/2}) \leq dp^{-1/2}$$

for a constant $d < \infty$. Now the minimaxity of $\hat{\mu}_M$ implies that

$$\sup_{\mu\in\Theta}\inf_h Ep^{-1}\sum_{i=1}^p(\hat{\mu}_{i,M}(h) - \mu_i)^2 = \inf_{\hat{\mu}\in\mathcal{L}}\sup_{\gamma'\tilde{B}\gamma\leq\rho} Ep^{-1}\sum_{i=1}^p(\hat{\mu}_i - \mu_i)^2 =: \nu_p^2$$

where $\mathcal{L}$ is the class of all linear estimators. Then, as $p \to \infty$, $p\nu_p^2 \geq \inf_h E\sum_{i=1}^p(\hat{\mu}_{i,M}(h)-\mu_i)^2 \to \infty$ and hence

$$\sup_{\mu\in\Theta} Ep^{-1}\sum_{i=1}^{p}(\hat{\mu}_{i,M}(\hat{h})-\mu_i)^2 = \inf_{\hat{\mu}\in\mathcal{L}}\sup_{\mu\in\Theta} Ep^{-1}\sum_{i=1}^{p}(\hat{\mu}_i-\mu_i)^2(1+o(1)) \tag{26}$$

for any $0 < \rho^* < \infty$. The minimax risk is attained by the adaptive minimax estimator where h is chosen through Mallows' C_L. This is for σ^2 known. However, it continues to hold if $\hat{\sigma}^2$ satisfies Kneip's eq.(6.2) which is trivial if $\hat{\sigma}^2 = \hat{\sigma}^2_{LS}$ because mean and variance are estimated from independent data. Since we need consistent variance estimation, we also need $n - p \to \infty$ as well as $p \to \infty$. □

Acknowledgements

This work was supported by a postdoctoral fellowship from the Research council of Norway. Some of the research was carried out while the author was a doctoral student at the University of California, Berkeley, supported by grant 411.92/001 from the Norwegian Research Council and by grant DMS 92-24868 from the National Science Foundation. The author would like to thank Professor Rudy Beran for many helpful discussions.

References

Becker, R.A.,Chambers, J.M. and Wilks, A.R. (1988). *The new S language.* Bell Telephone Laboratories, Murray Hill, New Jersey.

Beran, R. (1995). Stein confidence sets and the bootstrap. *Statistica Sinica* **5**, 109-127.

Breiman, L. (1995). Better subset regression using the nonnegative garotte. *Technometrics* **37**, 373-384.

Buja, A.,Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Carter, C.K., Eagleson, G.K. and Silverman, B.W. (1992). A comparison of the Reinsch and Speckman splines. *Biometrika* **79**, 81-91.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109-148.

George, E.I. and Oman, S.D. (1996). Multiple-shrinkage principal component regression. *The Statistician* **45**, 111-124.

Golub, G.,Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Hald, A. (1952). *Statistical theory with engineering applications.* Wiley, New York.

Heckman, N.E. and Woodroofe, M. (1991). Minimax Bayes estimation in nonparametric regression. *Ann. Statist.* **19**, 2003-2014.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

James, W. and Stein, C.M. (1961). Estimating with quadratic loss. *Proc. 4th Berkeley Symp. Math. Statist. Probab.* **1**, 361-380. University of California Press.

Kneip, A. (1994). Ordered linear smoothers. *Ann. Statist.* **22**, 835-866.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14**, 1101-1112.

Mallows, C.L. (1973). Some comments on C_P. *Technometrics* **15**, 661-675.

Massy, W.F. (1965). Principal component regression in explanatory statistical research. *J. Amer. Statist. Assos.* **60**, 234-256.

Næs, T. (1985). Multivariate calibration when the error covariance matrix is structured. *Technometrics* **27**, 301-311.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13**, 984-997.

Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian white noise. *Problems Inform. Transmission* **16**, 120-133.

Pilz, J. (1986). Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plann. Inference* **13**, 297-318.

Sclove, S.L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc.* **63**, 596-606.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970-983.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Weisberg, S. (1980). *Applied Linear Regression.* Wiley, New York.

Percentile	$\hat{\beta}_M$	$\hat{\beta}_R$	$\hat{\beta}_{PC}$
10 %	0.429	0.427	0.615
25 %	0.652	0.688	0.802
50 %	0.871	0.865	0.993
75 %	1.040	1.047	1.158
90 %	1.167	1.355	2.057
95 %	1.267	2.174	3.050
average	0.847	1.005	1.235
sd	0.312	0.971	0.979

Table 1: APSE ratio comparison for cement data

Percentile	$\hat{\beta}_M$	$\hat{\beta}_R$	$\hat{\beta}_{PC}$
10 %	0.718	0.732	0.836
25 %	0.786	0.801	0.956
50 %	0.863	0.882	1.000
75 %	0.968	0.972	1.078
90 %	1.110	1.116	1.272
95 %	1.252	1.233	1.469
average	0.899	0.906	1.040
sd	0.196	0.172	0.214

Table 2: APSE ratio comparison for car price data

Percentile	$\hat{\beta}_M$	$\hat{\beta}_R$	$\hat{\beta}_{PC}$
10 %	0.239	0.244	0.224
25 %	0.353	0.352	0.325
50 %	0.539	0.524	0.488
75 %	0.749	0.728	0.686
90 %	0.932	0.873	0.845
95 %	1.011	0.950	0.920
average	0.568	0.545	0.517
sd	0.269	0.237	0.240

Table 3: APSE ratio comparison for highway accident data

Percentile	$\hat{\beta}_M$	$\hat{\beta}_R$	$\hat{\beta}_{PC}$
10 %	0.448	0.471	0.542
25 %	0.610	0.622	0.711
50 %	0.784	0.791	0.919
75 %	0.935	0.956	1.009
90 %	1.070	1.129	1.159
95 %	1.156	1.243	1.259
average	0.771	0.802	0.880
sd	0.245	0.296	0.260

Table 4: APSE ratio comparison for the fish data

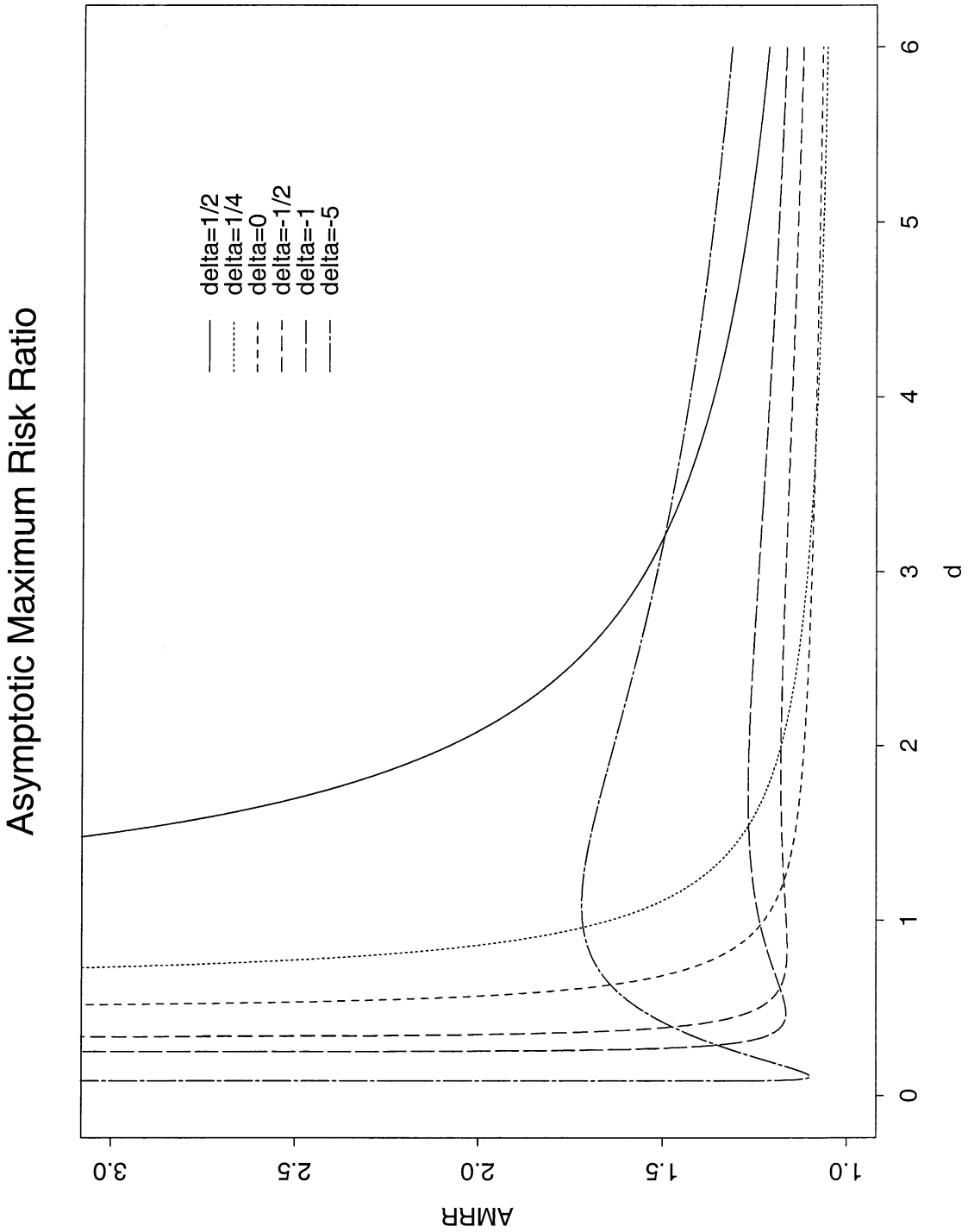

Figure 1: Asymptotic risk ratio for power ridge regression vs. minimax estimator.

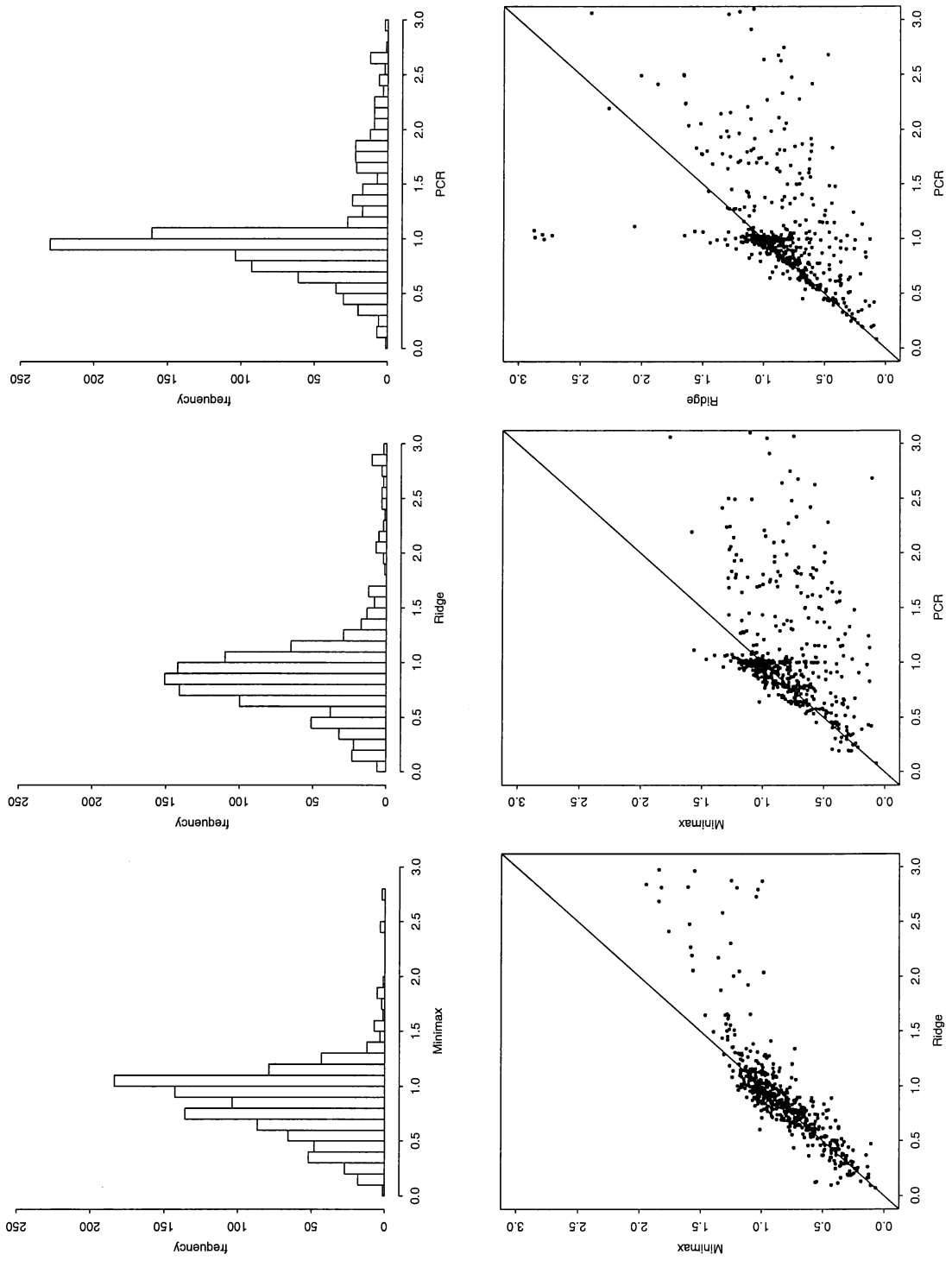

Figure 2: Average prediction squared error ratios for cement data relative to OLS.

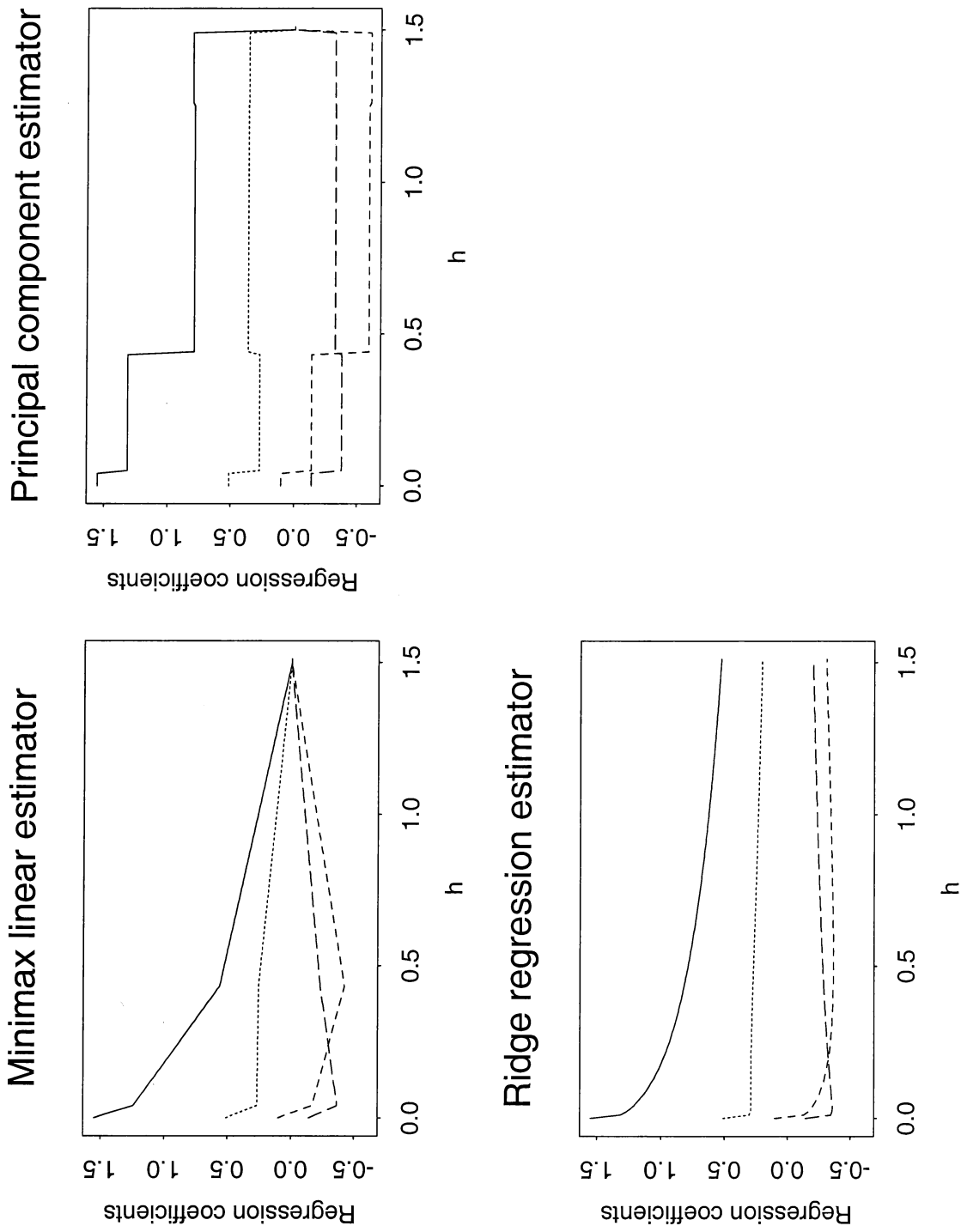

Figure 3: "Ridge trace" for $\hat{\beta}_M$, $\hat{\beta}_{PC}$ and $\hat{\beta}_R$ for cement data.

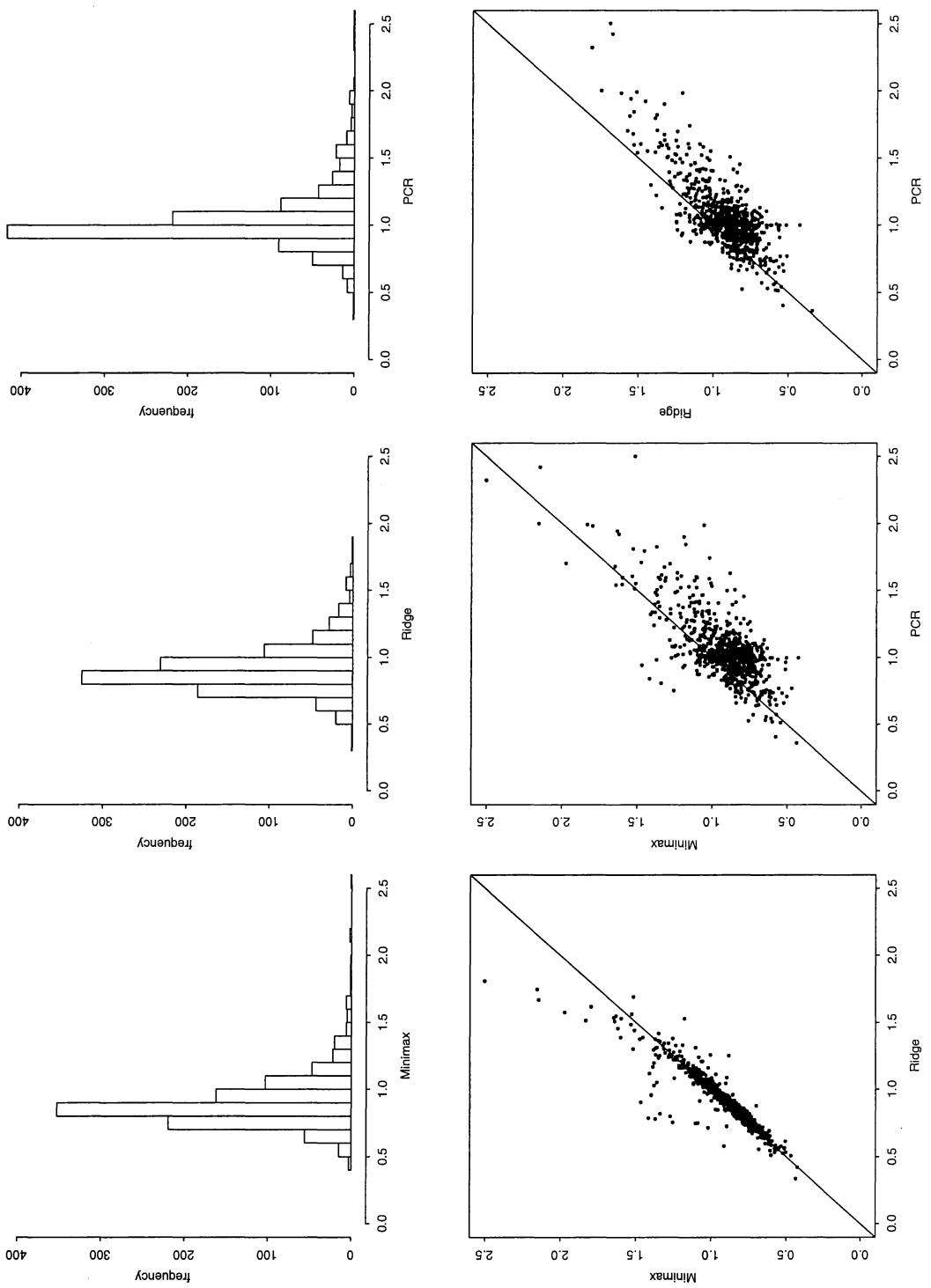

Figure 4: Average prediction squared error ratios for car price data relative to OLS.

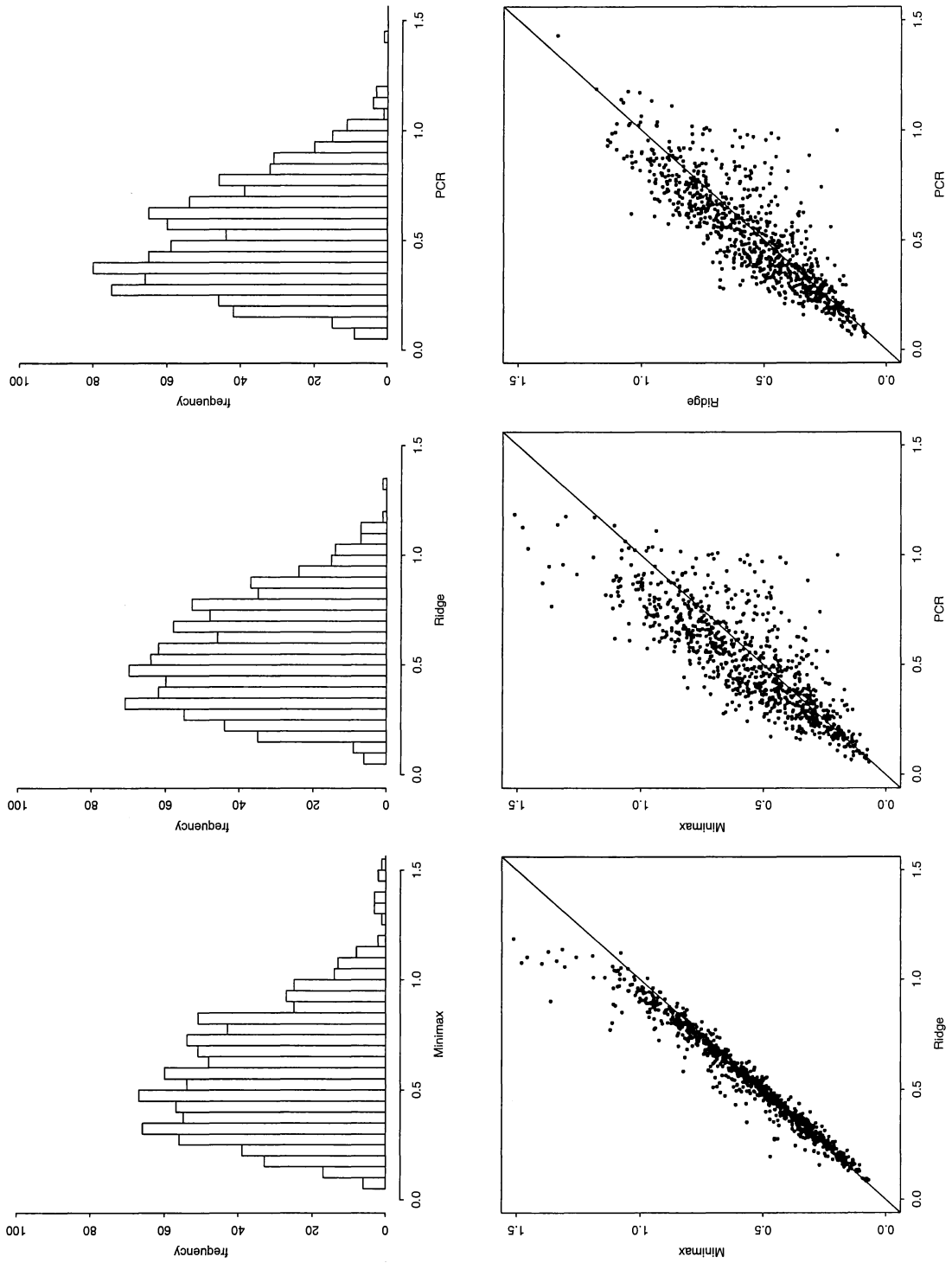

Figure 5: Average prediction squared error ratios for highway accident data relative to OLS.

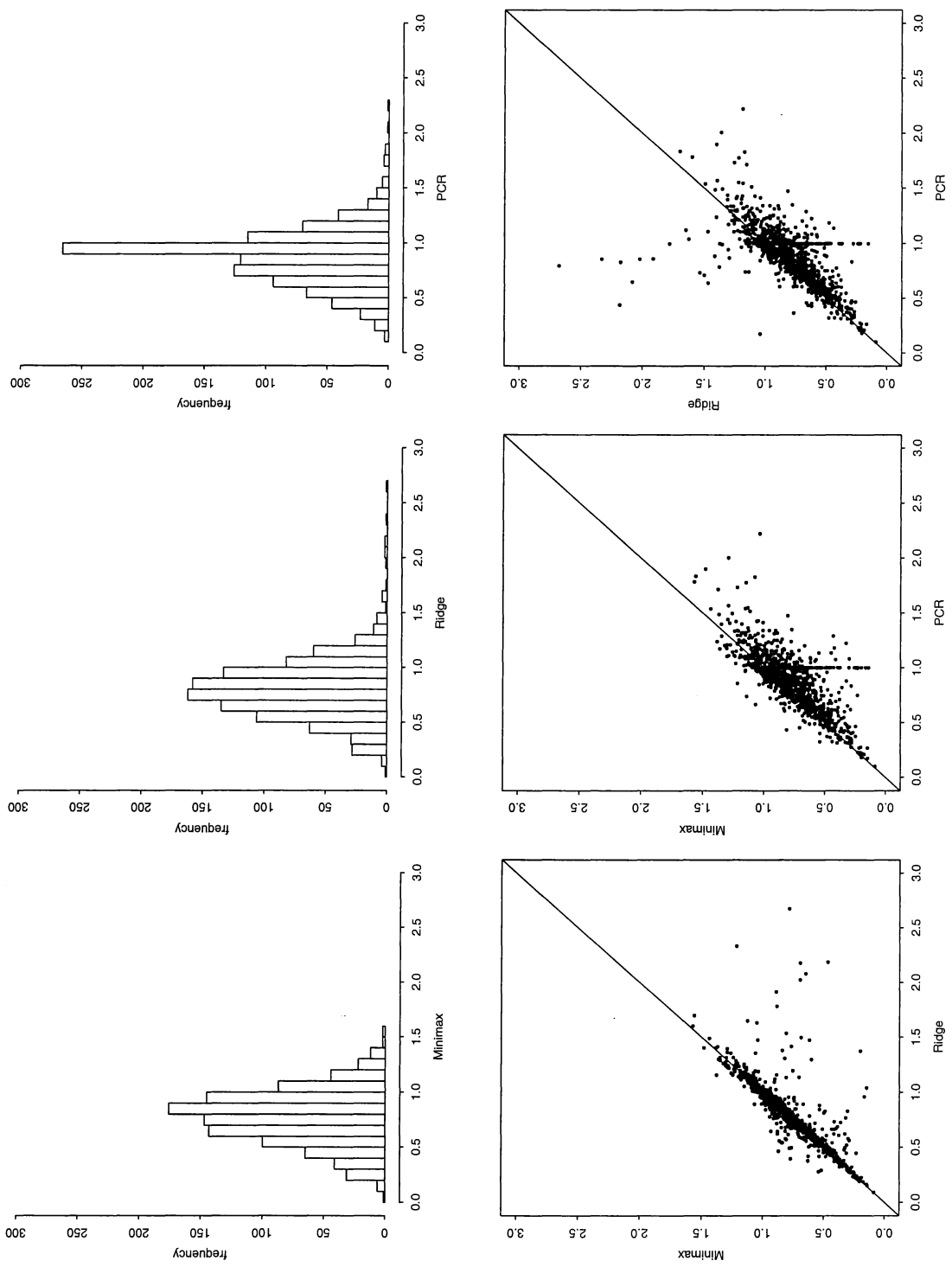

Figure 6: Average prediction squared error ratios for the fish data relative to OLS.